Usage Metrics for Research Data: Shortcomings, Mitigations and Research Questions

Data Usage Metrics Working Group
Research Data Alliance

Abstract—The fair assessment of research data is a hard task, both for technical and social reasons. Usage metrics are a central pillar in solving these problems. However, usage statistics are not without shortcomings, which are presented and classified in this paper; technical and organizational mitigations exist or are in development, but there are still issues demanding further research. A list of such open challenges is another contribution of this paper. Facing these challenges will contribute to open and transparent scientometric research and fair assessment of digital scholarly output.

Index Terms—Research data, scientometrics, bibliometrics, usage metrics

I. INTRODUCTION

Research data are a key ingredient in evidence-based methods of modern research. Their importance is growing, as is the amount of research data available for researchers. These developments come with social and technical challenges; one of these challenges concerns the assessment of research data: With both the number and the size of data growing exponentially, how can we assess the quality, impact, or relevance of research data?

One approach to answer this question is based on usage metrics. Usage metrics of research data are a special case of event-based metrics for research data, which are metrics summing up interactions over (possibly distributed) sources, such as publications, logs of providers of research data services, or social media platforms. Other examples for event-based metrics are citations or likes on social media platforms. While these two latter examples are based on "public" interactions, usage statistics are "private" in the sense that the interaction with the dataset is not on public record. Examples are views and downloads of research data.

All event-based metrics can be taken as indicators for quality, impact, or relevance of a dataset. In this case several shortcomings of these metrics need to be taken into account, of which some are common to impact or popularity metrics in general, but some are also specific to the context at hand, namely research.

During the lifespan of the "Data Usage Metric" working group of the Research Data Alliance (RDA)¹, these shortcomings of usage statistics have been collected and discussed, most prominently during the 13th RDA plenary in Philadelphia (April 2019), the 17th ISSI conference in Rome (September 2019), and the 14th RDA plenary in Helsinki (October 2019). Main source for the discussion was the Code Of Practice for research data usage metrics” (COP, [1]), which has been adopted by a number of large research data repositories and data aggregators, such as DataOne [2], Dryad [3], and Zenodo [4]. This paper summarizes the results of these discussions. The contributions are as follows:

- A comprehensive list of shortcomings
- A conceptual framework to classify the shortcomings
- An evaluation of impact of each shortcoming as a first step towards risk assessment
- A mapping of each shortcoming to the role(s) responsible to mitigate them.
- A report on the technical status quo of the mitigations
- A list of open research questions considering all of the aforementioned aspects

The remainder of paper is structured as follows: Section II introduces the definition of basic concepts and presents a selection of related work. In Section III the shortcomings will be listed and defined, and a scheme for classification is proposed. The following Section IV presents a partial risk assessment and outlines mitigation strategies. Section V presents open research questions. The paper is concluded by Section VI which gives an outlook over the next planned activities.

II. BACKGROUND

A. Definitions

- A metric is a mapping of a dataset to the real numbers.
- The value the metric maps the dataset onto is called the score of the dataset.
- A transaction denotes the event during which an interaction with the dataset took place (download, view, citation, etc.).

B. Related work

[5] discusses heterogeneity, data quality and dependencies as challenges for altmetrics. Many aspects discussed in our paper are also raised by the authors (e.g. correlation, versioning, timeliness, normalization, bandwagon effect), she is considering the shortcomings from the scientometrician’s perspective.

[6] is an opinion piece on shortcomings of metrics in the context of assessing the impact of research. It is not limited to a certain type of metric. Beside the general scepticism with regard to any quantitative approach to assess research the paper includes a detailed discussion of citation-based metrics (which can also be classified as event-based metrics.). [7]

¹https://rd-alliance.org/groups/data-usage-metrics-wg
shares a sceptical stance of (alt)metrics, but points out possible beneficial use cases (faster results compared to citations and measuring societal relevance).

III. Shortcomings

The presented shortcomings were collected during the lifespan of the “Usage Data Metrics” working group of the Research Data Alliance. During the meetings (RDA plenaries, video conferences) and in outreach activities (participation at the ISSI in Rome), they were discussed and refined. The section starts with a characterization of the shortcomings, continues with a categorization proposal and concludes with a tabular mapping of the shortcomings to the categories.

A. List of Shortcomings

The following distinct shortcomings were brought up in this or a similar form:

1) **Interaction Counting**: The score of a dataset is too high, because it counts several interactions by one user (like double-clicks or sessions) as single transactions, instead of one.

2) **Same Name Duplication**: The score of a dataset is too low, because it does not include all transactions since it is (identically) published on several repositories.

3) **Different Name Duplication**: The score of a dataset is too low, because it was (re-)published under a different identifier with the same content, and these scores have not been merged.

4) **Good bot**: The score of a dataset is too high, because requests from bots (crawlers, spiders, etc.) are counted equally to “human” or “normal” requests. This shortcoming only includes behaviour that is unintentionally harmful. Evidence for these types of bots is discussed in [8] with regard to preprints and tweets, which is a similar setup than usage metrics.

5) **Forgery (Bad bot)**: The score of a dataset is too high, because the number of investigations and requests is artificially increased, probably with the help of machines and circumventing the technical countermeasures in place. This shortcoming only includes intentionally malevolent behaviour. As for the example of good bots, [8] discusses evidence and recommendations for a similar setup.

6) **Versioning**: The score of a dataset is too low, because its predecessors or successors are not accounted for. The paradigmatic case for this shortcoming is the existence of preprints and postprints for a publication (analogue problem with altmetrics: [5]).

7) **Granularity**: The score of a dataset is too high/low, depending on the granularity in which it is made accessible. If the dataset is published as one item, there is also only one transaction; if it is split up into components, there are several transactions. Heterogeneous datasets pose special granularity problems, e.g. when code and data are published in one package, it is unclear whether both or only one is used (and in this case: which one).

8) **Aggregation**: The score of a dataset is too high, because it is published both as a composite and component-wise, and the transactions are aggregated and counted both on the level of the components and on the level of the dataset. In contrast to the granularity shortcoming, the problem is not the resolution in which the data are published, but that one transaction is counted twice.

9) **Fancy usage**: The score of a dataset is too low, since a usage context that differs from the canonical usage context is not counted/reportable or countable/reportable. Canonical usage includes views (investigations in the COP) and downloads (requests in the COP). An example are streamed data: these are used differently than static data, their access pattern typically does not allow for easy discretizations comparable to static data publications.

10) **Timeliness**: The scores for datasets x and y are not comparable, because they were released at two distant points in time. Timeliness is especially a problem for usage statistics that are only recently collected, (identical problem with altmetrics: [9], [5]).

11) **Normalization**: The scores for datasets x and y are not comparable, because the number of people typically interested in datasets like x is bigger/smaller than the according number for y. The problem is discussed for publications in [10]. Different altmetrics scores are compared across disciplines in [11], but no proposals to normalize them for the sake of comparability are formulated.

12) **Bandwagon Effect**: You cannot use usage metrics to definitely determine impact/quality/relevance of the scored dataset, because already popular datasets are more likely to get increased usage scores [12], [13] (also called ”Matthew effect in [5], [14]). Comparisons between datasets are more likely to be skewed, the bigger the distance between the scores is.

13) **Trust**: You cannot use usage metrics when arbitrary repositories report their scores, because there is no reason to trust the reporting repository with regard to the numbers reported and with regard to the correct implementation of the standard.

14) **Correlation**: You cannot use usage metrics to determine impact/quality/relevance of the dataset, because usage metrics do not necessarily correlate with impact/quality/relevance: A ”bad example” could easily get high access numbers (see [15] for evidence), if creating sensations is incentivized the quality could suffer in the long run [7]. Zero or very low access numbers do not indicate low quality (a ”good” dataset could have not been promoted).

To be discussed:

- **Coverage**: The score of a dataset is too low, because

---

2 We found no methodological discussion of the different and often classified view counting mechanisms of big video platforms, such as youtube or vimeo

B. Classes of Shortcomings

As can be seen from the previous subsection, critique of usage metrics commonly follows one of the following patterns:

- The score of a dataset is too low/high, . . .
- The score of datasets x and y are not comparable, . . .
- You cannot use usage metrics in context z, . . .

This suggests to categorize the shortcoming of usage metrics by the direction of the critique:

- **validity**: the score of a dataset is invalid (replace with uncertainty?)
- **comparability**: scores of datasets are not comparable
- **contextuality**: the scores cannot be used in a certain context

Another category captures the role(s) most suited to design mitigation strategies for a shortcoming. The following generic roles have a specific perspective on the creation, aggregation and usage of usage metrics:

- **Repository Providers**: The role responsible for hosting the platform on which the transactions occur.\(^3\) Examples are Zenodo and Dryad. In Table I we did not count adhering to the standard as a mitigation strategy.
- **Standard Issuers**: The role responsible to define standards as [1].
- **Metric Aggregators**: The role responsible to collect the data from the RPs; examples are DataCite and DataOne
- **Metric Users**: The role that uses the (aggregated) metric for a certain purpose; an example are decision makers, who are assigning resources based on metrics.

As a last category, code of practice, we classify the shortcomings whether they are handled by the COP [1]. Since this is a matter of degree the possible values are entirely, if the COP handles the shortcoming exhaustive, partially if it handles some of the aspects of the shortcomings, and no, if the COP does not provide mitigations/countermeasures.

The mapping of classes to the shortcomings listed in the previous subsection can be seen in Table I.

C. Advantages and Strengths

In comparison to other event-based metrics, usage metrics have also advantages:

- The services are not provided by a quasi-monopolistic (such as youtube for videos or twitter for micro blogs), which allows to collaboratively define standards to ensure fairness and comparability.
- Shared with "traditional" social metrics (altmetrics) is the capability to capture interest much faster than more formal event-based metrics such as citations.

Other metrics which are not based on recorded events across "platforms" include data benchmarks. Data benchmarks are often very specific to a certain technology and expensive to implement, both shortcomings are not shared by event-based metrics (including usage metrics).

IV. RISK ASSESSMENT AND MITIGATION STRATEGIES

This section assesses the risk of each shortcoming and presents the state of the art of possible mitigation strategies. Risk is understood as a combination of the probability of occurrence and the impact of the potential damage. While the available data are too sparse for an estimation of probabilities of occurrence, the expected impact can be thoroughly discussed in this paper:

- **Minimal damage** means that the shortcoming has a negligible effect on assessments.
- **Medium damage** means that the shortcoming introduces a detectable bias, when the metrics are used for assessment, which could be corrected programmatically.
- **Large damage** means that the shortcoming has the potential to make a fair assessment impossible.

It is yet unknown, which role usage metrics will play in the assessment of the digital output of the sciences and humanities, but it is to be expected that if there is an incentive for higher scores, stakeholders will take advantage of any bias, skewness or shortcoming available [17].

Mitigations are discussed in the context of the roles responsible to implement them.

TODO: Does the standard allow revocation of already reported metrics?

A. Interaction Counting

Variation interaction counting is of large impact. The COP standardizes the counting procedure with regard to double-clicks and sessions (7.2/7.3). This is clearly an advantage over aggregators of altmetrics (social media metrics), where aggregation is unspecified and reported numbers vary across aggregators [18].

B. Same Name Duplication

Missing metrics due to this type of duplication is of medium impact, since the merging of two datasets with the same identifier is a programmatic countermeasure. The COP demands persistent and globally unique identifiers for datasets, which makes this shortcoming efficiently mitigateable by metrics aggregators. It would be beneficial to make statistics about the merging process publicly available, e.g. which sources reported which amount of transactions.

C. Different Name Duplication

Different name duplication is of large impact. The main difference between this shortcoming and same name duplication is the absence of an easy way to determine whether two datasets are equal. In a future in which more platforms take advantage of liberal licensing of some datasets this might become a bigger issue than it seems to be today. Even more so, if different repositories offer services of varying quality: repository A might curate a dataset x (which has

\(^3\)[1] lists repository and data repository as possible values.
TABLE I

<table>
<thead>
<tr>
<th>#</th>
<th>Shortcoming</th>
<th>Direction</th>
<th>Repository Provider</th>
<th>Standard Issuer</th>
<th>Metrics Aggregator</th>
<th>Metrics User</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Interaction Counting</td>
<td>validity</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>entirely</td>
</tr>
<tr>
<td>2</td>
<td>Same Name Duplication</td>
<td>validity</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>entirely</td>
</tr>
<tr>
<td>3</td>
<td>Different Name Duplication</td>
<td>validity</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>no</td>
</tr>
<tr>
<td>4</td>
<td>Good bot</td>
<td>validity</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>no</td>
</tr>
<tr>
<td>5</td>
<td>Forgery (bad bot)</td>
<td>validity</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>no</td>
</tr>
<tr>
<td>6</td>
<td>Versioning</td>
<td>validity</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>no</td>
</tr>
<tr>
<td>7</td>
<td>Granularity</td>
<td>validity</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>no</td>
</tr>
<tr>
<td>8</td>
<td>Aggregation</td>
<td>validity</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>no</td>
</tr>
<tr>
<td>9</td>
<td>Fancy usage</td>
<td>validity</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>no</td>
</tr>
<tr>
<td>10</td>
<td>Timeliness</td>
<td>comparability</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>partially</td>
</tr>
<tr>
<td>11</td>
<td>Normalization</td>
<td>comparability</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>no</td>
</tr>
<tr>
<td>12</td>
<td>Bandwagon</td>
<td>contextuality</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>no</td>
</tr>
<tr>
<td>13</td>
<td>Trust</td>
<td>contextuality</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>no</td>
</tr>
<tr>
<td>14</td>
<td>Correlation</td>
<td>contextuality</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>no</td>
</tr>
</tbody>
</table>

several versions over time of both data and metadata), while repository B only publishes content once and never applies any curational tasks on it. The COP prescribes unique reporting of datasets published in more than one repository (7.4), but this is ineffective in cases, in which the repositories are not provided under the same institutional roof.

The only role who can effectively mitigate this shortcoming is the repository provider on the one hand repository providers could allow submitters to specify already existing identifiers (Zenodo allows that); on the other hand repository providers should update the usage agreements to ban re-publishing of already published content with a differing identifier than the original one.

D. Good Bots

This shortcoming is of medium impact. The COP offers the possibility to differentiate between regular usage and usage by a machine based on the user agent field of the http request. It is furthermore differentiated between crawlers and bots from search engines, and scripted and automated access. Only the latter should be counted. The maintenance of a blacklist of user-agents to be dropped by the counters of the repository providers is maintained by the issuers of the standard.

E. Forgery (Bad Bots)

This shortcoming is of large impact. Automated attacks are known to be distributed and of variance in the http requests sent, which makes their detection and therefore countermeasures difficult. Possible countermeasures on the side of repository providers include partial reporting (if the malicious accesses can be narrowed down to a specific window of time), anomaly detection (a dataset getting anormal attention). A standardized way to revoke already reported metrics would also be an effective countermeasure to decrease the motivation of the attackers. Another challenge lies in the "private" nature of usage statistics: while other event-based metrics allow the quantification of self-promotions (citing or liking own content), there is no feasible way to do that for usage statistics.

F. Versioning

This shortcoming is of medium impact. The COP makes the reporting of a version specifier mandatory, if applicable. Transcation are furthermore reported for each version and as a cumulative total across versions.

G. Granularity

This shortcoming is of large impact. Since there is an incentive to publish research data in the highest resolution of granularity, it is possible that publishing strategies of repository providers converges to such a practice. The COP allows the reporting of the download volume, which introduces a reciprocal incentive, but this is not without its own problems, e.g. a bias for large datasets. There is in general no mitigation available for this shortcoming.

H. Aggregation

This shortcoming is of medium impact. It is mitigated by the COP: If a user requests a component of a dataset it is only counted as one unique request. If the composite and its components is published as single entities, their counts must be independent with regard to the COP.

I. Fancy Usage

This shortcoming is of medium impact. The COP states that it does not specify how usage metrics for "dynamically changing datasets" should be reported. Whether or not data streams (like video streams, audio streams, sensory data) fall under this category is not explicitly stated in the COP. Steps necessary to include fancy usage include the definition of appropriate transaction types, such as (partial) consumption of video or audio streams or recording of sensory streams. Unless such a conceptual framework is available, metrics users should refrain from comparisons of metrics for such "dynamic" content with the scores reported COP-compliant.

J. Timliness

This shortcoming is of large impact. The COP circumvents this problem by making the reporting of the publication year mandatory. This allows to normalize the number of transaction...
with regard to the age of the dataset. The problem persists, if a yearly resolution is insufficient for a temporal contextualisation of the scores. If instead of the year, the date of the publication would be made mandatory, a finer resolution would be possible. Unless a finer resolution is available, metrics users should make comparisons with care, if the datasets compared are both relatively young (although they have been published in the same year, they could be almost one year apart).

K. Normalization

This shortcoming is of large impact. It is necessary to retrieve organizational and technical background when comparing scores of two data sets. This is the responsibility of metrics users. Possible contexts to normalize against include but are not limited to:

- discipline of research
- formats and encodings
- size

L. Bandwagon Effect

This shortcoming is of large impact. For normalization procedures and analyses based on non-robust statistics (such as the arithmetic mean) outliers have the potential to bias the resulting values. Metrics users should discuss the effect of this shortcoming in their decisions or conclusions. The threat to overstress outliers can be mitigated by contextualizing scores with data benchmarks.

M. Trust

This shortcoming is of minimal impact. The COP states the possibility of audits without specifying what institution would be responsible for such an endeavor. Certified audits are a good countermeasure, if they are honored by metrics aggregator, which means that the possibility exist to select only those scores which were aggregated over repository providers which were audited. Metrics aggregators could themselves be audited.

N. Correlation

This shortcoming is of large impact. As with bandwagon effects the only available countermeasure is to contextualize the scores with other event-based metrics (citations, altmetrics) or data benchmarks. Metrics users should consider a discussion of the context in order to mitigate this shortcoming.

V. OPEN RESEARCH QUESTIONS

- Indicators for shortcomings: How can we measure these?
- Related to previous question: Success criteria for mitigations
- Usefulness as impact/quality/relevance indicator?
- Relation to other metrics (altmetric, citations, data benchmarks): Better, worse, complimentary?

Indicators do they occur? If so clustered in some disciplines?

VI. CONCLUSION & OUTLOOK

- RDA activities
- Make data count project

ACKNOWLEDGEMENTS

Remember to thank:

• 1
• 2

REFERENCES


Thursday 31st October, 2019 16:55